Методология интеллектуального анализа данных: перспективы применения в исследовании ценностных ориентаций студенчества

О.Н. Кислова, Л.Г. Сокурянская

В статье рассматриваются возможности и перспективы использования интеллектуального анализа данных (ИАД) в социологии. Раскрывается специфика ИАД, его технологии, выявляются типы знания, которые могут быть получены с помощью ИАД (поверхностное, многомерное, скрытое, глубокое). На примере исследования ценностных ориентаций приводятся методы, использование которых обеспечивает получение отмеченных типов знания. Очерчиваются перспективы последующего использования ИАД в социологических исследованиях.

У статті розглядаються можливості та перспективи використання інтелектуального аналізу даних (ІАД) у соціології. Розкривається специфіка ІАД, його технології, виокремлюються типи знання, що можуть бути отримані за допомогою ІАД (поверхневе, багатомірне, приховане, глибоке). На прикладі дослідження ціннісних орієнтацій наводяться методи, застосування яких призводить до отримання зазначених типів знання. Окреслюються перспективи подальшого використання ІАД у соціологічних дослідженнях.

Ключові слова: інтелектуальний аналіз даних, знання, ціннісні орієнтації.

У артыкуле разглядаюцца магчымасці і перспектывы выкарыстання інтэлектуальнага аналізу дадзеных (ІАД) у сацыялогіі. Раскрываецца спецыфіка ІАД, яго тэхналогіі, выяўляюцца тыпы ведаў, якія могуць быць атрыманы з дапамогай ІАД (павярхоўныя, шматмерныя, схаваныя, глыбокія. На прыкладзе даследавання каштоўнасных арыентацый паказваюцца метады, выкарыстанне якіх прыводзіць да атрымання адзначаных тыпаў ведаў. Акрэсліваюцца перспектывы наступнага выкарыстання ІАД ў сацыялагічных даследаваннях.

Ключавыя словы: інтэлектуальны аналіз дадзеных, веданне, каштоўнасныя арыентацыі.

Opportunities and prospects of using intelligent analysis of sociological data (IAD) are examined in the article. The specificity of IAD, its technologies are studied, different types of knowledge that can be received using IAD (superficial, multidimensional, latent, deep) are described. Taking as an example the results of value orientations studies the authors describe methods that can be applied to gain the above-mentioned types of knowledge. The prospects of further use of IAD in sociological researches are defined.

Keywords: data mining, knowledge, value orientations.

В настоящее время по самым грубым оценкам количество данных удваивается каждый год, а количество значимой информации соответственно быстро уменьшается. Это обусловливает непрерывный поиск возможностей эффективного извлечения полезной информации из необозримого океана данных. Исследования в области искусственного интеллекта позволили разработать методологию интеллектуального анализа данных (ИАД), которая не зависит от конкретной предметной области и может с успехом применяться для решения широкого спектра задач в самых разных научных и практических сферах.

Интенсивное применение интеллектуального анализа данных осуществляется благодаря наличию рабочих инструментов (пакетов программ), реализующих разнообразные методы ИАД. Эксперты считают, что в ближайшее десятилетие интеллектуальный анализ данных и его ядро – Data Mining – станут наиболее перспективными направлениями разработки программного обеспечения. По заявлению влиятельного издания MIT Technology Review, Data Mining — одна из десяти развивающихся технологий, которые «изменят мир»[1]. Потенциал ИАД велик, но есть и другая сторона. Уже сейчас наблюдается чрезмерное рекламирование программных реализаций методов ИАД, что связано с коммерциализацией результатов научной деятельности. Реклама ИАД часто вводит потенциальных пользователей в заблуждение, создает ошибочные представления, мешающие пониманию сути интеллектуального анализа, его возможностей и ограничений. Среди мифов об ИАД назовем следующие:

1) абсолютная новизна;

2) простота;

3) возможность найти решение исследовательской задачи благодаря применению «самого нового» метода или программного продукта ИАД.

Попробуем развеять эти мифы.

Во-первых, *новое направление* развития интеллектуальных систем различного назначения, в частности, *ИАД, нельзя назвать абсолютно новым*. Практически весь математический инструментарий интеллектуального анализа данных существует в рамках различных разде-

[1] Николаев А.Б., Фоминых И.Б. Интеллектуальный анализ и обработка данных: учеб. пособие. М., 2003; Data Mining: Дадим слово критикам [Электронный ресурс] Режим доступа: http://www.iso.ru/journal/articles/277.html

лов прикладной математики, статистики и кибернетики достаточно давно. Методы, используемые для интеллектуального анализа данных, применялись также и при традиционном построении количественных моделей. Так, например, нейронные сети, которые в настоящее время являются одним из наиболее востребованных интеллектуальных методов, были изобретены еще в 1940-х гг. Группа методов, называемых деревьями классификации, является по своей сути развитием регрессионных методов и использовалась специалистами в области общественных наук уже в 1960-х гг. Метод К-ближайшего соседа применяется более полувека.

Несостоятельность *второго мифа* проявляется в необходимости методично и вдумчиво интерпретировать информацию, полученную в результате применения методов ИАД. А этот процесс нельзя назвать простым. Кажущаяся простота ИАД объясняется появлением программных средств с удобным интерфейсом, которые позволяют легко применять сложный математический аппарат специалистам предметных областей, не являющихся профессиональными математиками и программистами. Алгоритмы для интеллектуального анализа данных часто являются сложными, однако их применение, благодаря появлению новых информационных средств, значительно упростилось. При этом следует обратить внимание на так называемые «простые» методы ИАД – визуализацию содержащихся в данных закономерностей в виде табличных и графических представлений, которые позволяют узнать немало интересного об исследуемом явлении, не применяя при этом сложных математических методов. Однако применение «простых» методов визуализации сопровождается трудоемкой интерпретацией результатов. Никакая интеллектуальная система не избавит аналитика от необходимости думать.

Наконец, *третий миф* развеивается сам собой в процессе исследования любого сложного социального явления. Попытки найти объяснения факторов многогранных феноменов на основе применения одного, даже самого нового метода анализа данных не дают полноценных результатов, поскольку каждый метод имеет ограничения и предназначен для анализа только вполне определенных аспектов изучаемого объекта. Интеллектуальный анализ данных – это процесс, последовательность действий, которую необходимо выполнить для извлечения из эмпирических данных нового знания (построения модели). Эта последовательность не описывает конкретный алгоритм или математический аппарат, не зависит от предметной области. Это набор разнообразных методов, выступающих в качестве атомарных операций, комбинируя которые, можно получить нетривиальные результаты. Подчеркнем, чтобы реализовать действительно интеллектуальную обработку данных и достигнуть значимого результата, необходимо иметь в своем распоряжении целый арсенал разнообразных методов анализа данных и программных инструментов их реализации.

Формирование интеллектуального анализа данных как нового научного направления происходит на основе обобщающих концептуальных положений. И кибернетика, и теория систем, и синергетика появились именно таким образом. Эти синтетические науки нашли широкое применение в разнообразных областях знания, в частности, в социологии. Именно это обусловливает *актуальность* рассмотрения и потенциала, и специфики ИАД в социологических исследованиях.

Поскольку под *интеллектуальным анализом* разные авторы понимают разное, представляется необходимым определить, что мы подразумеваем под этим понятием. *ИАД* – это некие механизмы, преобразовывающие данные эмпирического социологического исследования к высокоуровневым информационным объектам, более пригодным для анализа, чем исходная информация. Условно можно разделить эти механизмы анализа на две части – *методы визуализации данных* и *моделирование*. Благодаря первой группе методов социолог может обнаружить поверхностное или в лучшем случае многомерное знание. В контексте второй группы внимание акцентируется на получении модели для оценки влияния на анализируемый социальный феномен или процесс различных (как внешних, так и внутренних) факторов.

Особого внимания заслуживает рассмотрение специфики интеллектуального анализа социологических данных. Прежде всего отметим, что социологические эмпирические данные, т.е. данные, характеризующие конкретные социальные факты, могут представать перед исследователем в виде:

- чисел, характеризующих те или иные объекты;
- индикаторов определенных отношений между рассматриваемыми объектами;
- совокупности определенных высказываний;
- текстов документов;
- каким-либо способом зафиксированных результатов наблюдения за невербальным поведением каких-либо людей и т.п.[1]

Интеллектуальный анализ дает возможность работать со всеми видами социологических эмпирических данных. Единственное, что здесь стоит отметить – программное обеспечение для работы с качественными данными имеет свои особенности, привычные статистические пакеты в данном случае не могут применяться[2].

Интеллектуальный анализ данных позволяет находить *не только статистические закономерности, т.е. закономерности «в среднем», но и модели (закономерности), которые описывают, объясняют и прогнозируют редко встречающиеся явления.* Этим и определяется важность ИАД для социологии, поскольку социологи всегда стремились дополнить исследование средних тенденций изучением индивидуальных особенностей, что проявилось в признании эффективности анализа социальных феноменов одновременно количественными и качественными методами. Методология ИАД совмещает анализ средних тенденций и тенденций отклонения от средних (т.е. анализ редких, возможно, даже уникальных случаев) в рамках только количественного подхода. Аналогично, качественные исследования, проводимые с использованием ИАД, также интегрируют названные аспекты анализа данных.

Отметим, что Ю.Н. Толстова, обосновывая важность статистического подхода к анализу социологических данных, предлагает разделить все возможные социологические парадигмы на две группы по не совсем привычному основанию: *содержательные и методные*. Среди *методных* она, в свою очередь, выделяет *статистическую* и *системную парадигмы*. В соответствии с *системной парадигмой* изучение социальных объектов (социальных групп или отдельных индивидов) можно рассматривать как систему, придерживаясь

---

[1] См.: Татарова Г.Г. От постулатов эмпирической социологии к методологии анализа данных // Социология 4М. 1999. № 11; Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М., 2000.

[2] Поскольку социологические данные, получаемые качественными методами, чаще всего представляют собой тексты, то для их анализа методами ИАД можно использовать средства, которые получили название Text Mining. К ним относятся такие программные комплексы, как, например, IBM Intelligent Miner for Text, Oracle InterMedia Text, Megaputer Text-Analyst и др. (Ландэ Д. Глубинный анализ текстов – технология эффективного анализа текстовых данных. Добыча данных [Электронный ресурс] Режим доступа: http://www.visti.net/~dwl/art/dz/ Для анализа качественной социологической информации представляют интерес методологические и технологические разработки российских авторов (Гусакова С.М.; Михеенкова М.А., Финн В.К. О логических средствах автоматизированного анализа мнений // НТИ. Сер. 2. Информ. процессы и системы. 2001. № 5. С. 4-24; Данилова Е.Н., Климова С.Г., Михеенкова М.А. Возможности применения логико-комбинаторных методов для анализа социальной информации // Социология 4М. 1999. № 11. С. 142-160; Михеенкова М.А. Развитие ДСМ-метода автоматического порождения гипотез для его применения при анализе социологических данных типа «субъект-поведение»: автореф. дисс. ... канд. техн. наук / ВИНИТИ. 1998; Финн В.К. Об интеллектуальном анализе данных // Новости Искусственного интеллекта. 2004. № 3; Финн В.К., Михеенкова М.А., Бурковская Ж.И. О логических принципах анализа электорального поведения // НТИ. Сер. 2. Информ. процессы и системы. 2004 . № 8; Финн В.К. Об интеллектуальных системах типа ДСМ для наук о жизни и социального поведении // НТИ. Сер. 2, Информ. процессы и системы. 2002. № 6. С. 1-4). ДСМ метод является средством, которое можно рассматривать как высокоформализованный вариант качественного анализа социальных данных и уточнения моделей (Данилова Е.Н., Климова С.Г., Михеенкова М.А. Возможности применения логико-комбинаторных методов для анализа социальной информации // Социология 4М. 1999. № 11. С. 142-160). Интересным для социологов, которых привлекают качественные методы интеллектуального анализа, будет программа *синтаксического анализа*, которая доступна на сайте http://www.aot.ru/ (разработчики Л. Гершензон, Т. Кобзарева, Д. Панкратов, А.Сокирко, И. Ножов).

соответствующих системных принципов. Заметим, что названные парадигмы отнюдь не противоречат друг другу и могут эффективно использоваться одновременно, если ввести понятие *статистической системы*[1]. В этом контексте проявляются преимущества ИАД, позволяющего осуществить интеграцию поиска средних тенденций в более обширную программу выявления всех возможных знаний-закономерностей-моделей.

Новое знание – цель интеллектуального анализа данных, поэтому необходимо определить, что понимается под этим термином. *Формализации понятия «знание»* посвящено множество работ в области искусственного интеллекта[2]. Однако эти определения акцентируют внимание на способах разработки программных продуктов интеллектуального анализа. Для целей анализа данных социологического исследования более приемлемым будет следующее понимание данного термина. *Знание* в контексте методов ИАД означает отношения между элементами данных и шаблонами (паттернами, закономерностями), выявленными из данных. Таким образом, знание приравнивается к *модели, описывающей взаимосвязь исходных данных и найденных на их основе закономерностей.*

В технологии ИАД нужно различать четыре различных типа знания, которые могут быть извлечены из данных: поверхностное, многомерное, скрытое и глубокое знание[3].

1) *Поверхностное знание.* Это информация, которая может быть достаточно легко извлечена из массива данных, чаще всего в результате анализа одномерных и двумерных распределений.

2) *Многомерное знание* – информация, которую получают в результате применения различных методов статистического многомерного анализа, дающих возможность классифицировать, упорядочивать и структурировать исходные данные: от анализа многомерных распределений признаков до логлинейного анализа.

3) *Скрытое знание* представляет собой информацию, которая может быть получена разнообразными алгоритмами распознавания образов, например, кластерного анализа (распознавания образов без учителя) или дискриминантного анализа (распознавания образов с учителем); методами факторного анализа; многомерного шкалирования и др.

4) *Глубокое знание.* Это информация, которая хранится в базе данных, но может быть обнаружена только, если имеется ключ, который показывает нам, где смотреть. Ключом обычно являются теоретические представления об исследуемом явлении. Необходимо отметить, что указанное теорией направление поиска часто дает результаты, не прогнозируемые исследователем. Именно в этом и проявляется «новизна» получаемого знания. Последовательное применение различных методов совместно с анализом кажущихся расхождений в их результатах и соотнесение с социологическими теориями исследуемого феномена позволяет выявить этот тип знания.

Продемонстрируем, как были получены нами названные типы знания в контексте анализа динамики ценностных ориентаций студенческой молодежи.

Прежде всего отметим, что кафедра социологии и социологическая лаборатория Харьковского национального университета имени В.Н. Каразина обладают значительной базой данных, характеризующих субъективный мир студенчества (его интересы и потребности, ценностные ориентации и установки, мотивацию в различных сферах деятельности и т.д.). Поскольку исследования проблем жизнедеятельности будущих специалистов были начаты университетскими социологами еще в конце 1960-х гг. и не прекращаются до сегодняшнего дня, у нас есть уникальная возможность, сравнив их результаты, получить разнообразную

---

[1] См.: Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М., 2000.

[2] Вагин В.Н. Знание в интеллектуальных системах // Новости искусственного интеллекта, 2002. № 6 (54) [Электронный ресурс] Режим доступа: http://www.raai.org/about/persons/vagin/pages/vagin_zn.doc; Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. 2004. № 3.

[3] Николаев А.Б., Фоминых И.Б. Интеллектуальный анализ и обработка данных: учеб. пособие. М., 2003; Han J., Kamber M. Data Mining. Concept and Techniques. Morgan Kaufman Publishers, 2000.

эмпирическую информацию, в частности, данные о ценностных преференциях вузовской молодежи разных поколений.

Анализ одномерных и двумерных распределений материалов исследований конца 1960-х, 1970-х, 1980-х, 1990-х и начала 2000-х гг. позволил нам зафиксировать существенные изменения ценностного мира студенчества, получив то самое *поверхностное* знание, о котором мы писали выше. Но мы не могли ограничиться этим уровнем знания.

Целью следующего шага нашего изучения было определение направленности выявленной нами ценностной трансформации. Обратившись к анализу многомерных распределений, мы убедились в том, что ценностное сознание студенчества (особенно 1990-х и начала 2000-хгг.) изменяется в сторону *модернизации и постмодернизации*[1]. Осуществив эмпирическую типологизацию данных на основе социокультурных (цивилизационных) критериев отнесения ценностей к *традиционалистским, модернистским и постмодернистским*, используя *процедуру кластерного анализа*, мы получили пять групп (кластеров) респондентов, различающихся ориентациями на эти типы терминальных и инструментальных ценностей, а также реальным уровнем развития последних. Полученное нами (теперь уже *многомерное*) знание актуализировало выдвижение новых гипотез о дуализме как модернистского, так и постмодернистского ценностного дискурса молодежного (студенческого) сознания.

Переход от многомерного к обнаружению *скрытого* знания был осуществлен нами с помощью *многомерного шкалирования*[2]. Этот метод позволил нам свести ценностный континуум к двум шкалам: «прагматизм-идеализм» и «коммунализм-индивидуализм». Анализ корреляции построенных шкал с ценностными ориентациями студентов выделенных кластеров подтвердил нашу гипотезу о том, что модернизация ценностного сознания будущих специалистов осуществляется по оси «коммунализм-индивидуализм», а постмодернизация – по оси «прагматизм-идеализм». При этом положение кластера в пространстве построенных осей зависело от ориентации и уровня развития инструментальных ценностей (качеств личности) как важнейших средств реализации ценностей-целей (терминальных ценностей) представителей каждого кластера. Таким образом, выделенные кластеры получили название: *«модернисты-коммуналисты», «модернисты-индивидуалисты», «постмодернисты-прагматики», «постмодернисты-идеалисты» и «новые традиционалисты»*. Заметим, что последний кластер получил свое название благодаря тому, что представляющие его студенты, с одной стороны, исповедовали традиционные ценности, с другой – ориентировались на качества (инструментальные ценности) модернистской и даже постмодернистской направленности.

Для того чтобы понять, чем детерминируется амбивалентность ценностного сознания современного студенчества, т.е. получить, согласно представленной выше классификации, *глубокое* знание, мы совместили несколько различных методов, в том числе методы качественного анализа[3]. Тем не менее, это позволило нам разработать *мини-теорию ценностного поля личности,* содержащую, как представляется, определенный эвристический потенциал для объяснения как процессов ценностной динамики, так и бесконфликтности поливалентного ценностного дискурса личности. Кроме того, анализ качественной информации позволил нам существенно расширить свои знания о факторной обусловленности ценностных преференций студенчества.

Таким образом, с помощью интеллектуального анализа данных социолог-исследователь может получить подлинно *научный результат, новое знание* о состоянии явлений и процессов, которые он изучает, и определить новые возможные горизонты своих научных поисков.

---

[1] Сокурянская Л.Г., Кислова О.Н. Ценностная дифференциация украинского студенчества: кластерный анализ // Методологія, теорія та методи соціологічного аналізу сучасного суспільства. Збірник наукових праць. Харків, 2003. С. 534-539.

[2] Кислова О.Н., Сокурянская Л.Г. Многомерное шкалирование ценностных ориентаций студенческой молодежи // Вісник Харківського національного університету ім. В.Каразіна. 2005. № 652. С. 101-107.

[3] Мы имеем в виду материалы, полученные нами с помощью фокусированных групповых интервью, а также биографического метода.

Наши предыдущие публикации являются фрагментами ИАД, демонстрирующими возможности отдельных методов в исследовании ценностных ориентаций студентов[1].

В общем случае процесс интеллектуального анализа (т.е. поиска нового знания) состоит из следующих этапов[2]:

1) отбор данных (выбор признаков, которые, по предположению исследователей, являются значимыми для конкретного исследования, в нашем случае для анализа ценностных ориентаций современного студенчества);

2) предобработка данных или очистка (устранение неточностей, принятие решения о работе с неответами и т.д.),

3) трансформация (преобразование шкал для применения выбранных методов);

4) собственно извлечение знаний (Data Mining);

5) интерпретация результатов в контексте содержательных гипотез (см. рис. 1).

Главным элементом всего этого процесса являются методы Data Mining, позволяющие обнаруживать *новые* закономерности (шаблоны) и знания, являющиеся следствием интерпретации найденных закономерностей.

Можно утверждать, что Data Mining обозначает такой подход к анализу эмпирических данных, при котором исследователь готов к тому, что анализируемый феномен может оказаться слишком запутанным и не поддающимся точному анализу с помощью традиционных методов. Но, несмотря на это, пытается все же получить более полное представление о факторах, обусловливающих исследуемое явление и даже прогнозировать его дальнейшее развитие в различных контекстах, подходя к задаче с различных точек зрения, руководствуясь опытом и знанием предметной области, используя различные эвристические подходы. При этом социолог движется от самой простой модели (т.е. от уровня поверхностного знания) ко все более полным представлениям об анализируемом феномене (т.е. до уровня глубокого знания).

Таким образом, Data Mining подразумевает, что:

- при анализе следует использовать опыт специалистов предметной области;
- исследуемое явление необходимо рассматривать с разных точек зрения, что означает необходимость комплексного применения разнообразных методов;
- искать решение поставленной задачи следует поступательно: переходя от простых моделей ко все более сложным и точным;
- по прошествии времени, когда будут накоплены новые данные, целесообразно повторить цикл поиска нового знания.

В процессе Data Mining можно выделить три взаимосвязанных компонента[3]:

1) свободный поиск всевозможных закономерностей;

2) использование выявленных закономерностей для прогностического моделирования;

---

[1] Кислова О.Н. Интеллектуальный анализ данных: возможности и перспективы применения в социологических исследованиях // Методологія, теорія та практика соціологічного аналізу сучасного суспільства: Збірник наукових праць. Харків, 2005. С. 237-243; Кислова О.Н., Сокурянская Л.Г. Использование метода многомерного шкалирования в исследовании ценностей студенчества: процедура и результаты // Методологія, теорія та методи соціологічного аналізу сучасного суспільства: Збірник наукових праць. Харків, 2002. С. 543-546; Кислова О.Н., Сокурянская Л.Г. Многомерное шкалирование ценностных ориентаций студенческой молодежи // Вісник Харківського національного університету ім. В.Каразіна. 2005. № 652. С. 101-107; Сокурянская Л.Г., Кислова О.Н. Постмодернизация ценностного сознания современного студенчества: украинский и белорусский вариант // Методологія, теорія та методи соціологічного аналізу сучасного суспільства: Збірник наукових праць. Харків, 2004. С. 527-533; Сокурянская Л.Г., Кислова О.Н. Ценностная дифференциация украинского студенчества: кластерный анализ // Методологія, теорія та методи соціологічного аналізу сучасного суспільства: Збірник наукових праць. Харків, 2003. С. 534-539.

[2] Николаев А.Б., Фоминых И.Б. Интеллектуальный анализ и обработка данных: учеб. пособие. М., 2003; Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. 2004. № 3.

[3] Parsaye K. A Characterization of Data Mining Technologies and Processes // The Journal of Data Warehousing. 1998. № 1.

3) выявление наиболее нехарактерных шаблонов, т.е. анализ отклонений от средней тенденции, предназначенный для выявления и интерпретации аномалий в найденных закономерностях.

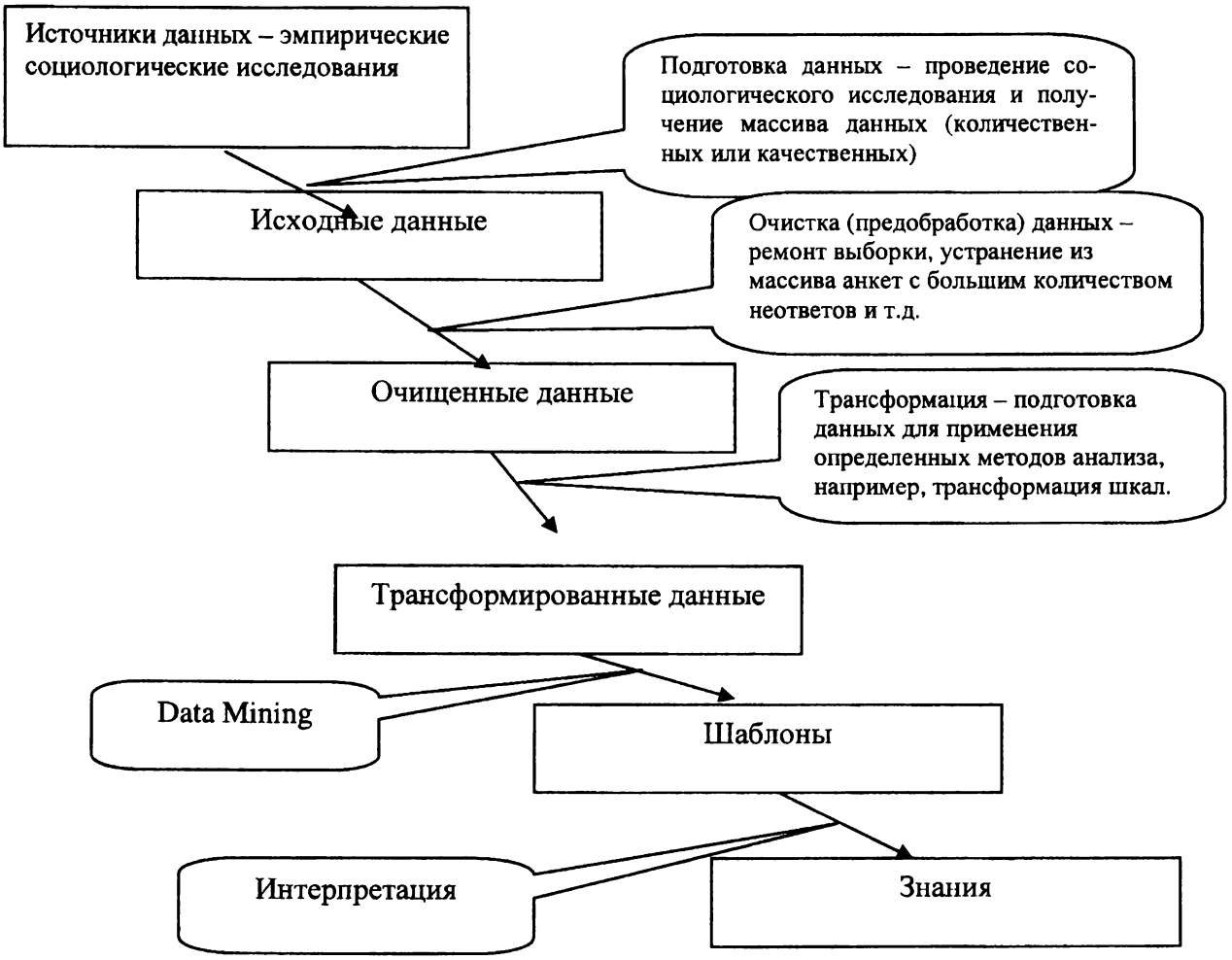

Рис. 1. Общая схема интеллектуального анализа данных социологического исследования

Таким образом, анализ возможностей применения ИАД в социологических исследованиях позволяет обозначить его перспективы в исследовании ценностных ориентаций студенчества:

1) применение методологии ИАД к анализу качественных данных, в частности, использование технологии Text Mining;

2) изучение отклонений ценностных ориентаций, которые могут стать доминирующими при определенных условиях развития социума;

3) прогнозирование трансформации ценностных ориентаций в условиях многовариативности и разновекторности развития нашего общества.